Library of Congress
September 11 Web Archive Project

Final Project Report

Submitted by

Research Foundation of the State University of New York

on behalf of the

SUNY Institute of Technology
WebArchivist.org

Steven M. Schneider
Principal Investigator

June, 2004

Table of Contents

Executive Summary

WebArchivist.org, a collaborative project at the University of Washington and the SUNY Institute of Technology, organized the September 11 Web Archive project in collaboration with the Library of Congress, the Internet Archive, and the Pew Internet & American Life Project. The Library issued a purchase order to the Research Foundation of the State University of New York (representing the SUNY Institute of Technology) for a processed and catalogued collection of URLs related to the September 11 terrorist attacks, including metadata about objects in the collection, and a Web-based interface to the collection.

This report is intended to serve as an overall summary of the project, and as a guide to future web archiving initiatives. The report summarizes the project expectations, the processes of the participants, and the project results. Most importantly, it identifies specific recommendations for future archiving efforts. Additionally, specific documentation of critical project components, including identified URLs for crawling and the structure of metadata databases, are provided in printed and electronic appendices.

In the conclusions and recommendations section, specific suggestions are provided with respect to each of the project components. These suggestions serve as the overall recommendations to the Library for future archiving projects:

- Develop a strategic approach, including implementation of anticipatory contracts to acquisition agencies, for identifying and archiving Web sites emerging in unanticipated, unstable and unpredictable Web spheres, and be prepared to implement this strategy as events required.

- In relatively small collections (fewer than 5,000 catalogued objects), do not rely on machine collection of metadata for catalog fields.

- Include detailed specifications for a Web application to serve as interface to archive in overall project plan.

Introduction

Project Background

The predecessor to this project began as an informal collaboration between researchers at the SUNY Institute of Technology and the University of Washington; a reference librarian at the Library of Congress; and an archivist at the Internet Archive. In a telephone conference call on September 12, 2001, the collaborators agreed to pool our resources and talents to identify and archive as many Web sites with content related to the terrorist attacks of the previous day as possible. Over the next few months, this informal project was formalized with a purchase order from the Library of Congress to the Research Foundation of the State University of New York (representing the SUNY Institute of Technology) for a processed and catalogued collection, including metadata about objects in the collection, and a Web-based interface to the collection.

Purpose of This Report

This report is intended to serve as an overall summary of the project as specified in the purchase order from the Library of Congress to the SUNY Research Foundation. While some background on the informal collaboration preceding this projected is provided as context, the emphasis is on the specific tasks and deliverables requested by the Library in the purchase order. The report summarizes the project expectations and the project results. It identifies specific recommendations for future efforts to archive rapidly emerging Web spheres. Specific documentation of critical project components, including the structure of metadata databases and the Web interface, are provided in electronic appendices.

Project Expectations

Acquisition and collection development

This project was designed to identify and catalog a set of Web sites selected from amongst those identified and archived between September 11, 2001 and November 30, 2001 by the Internet Archive. Under contract with the Library of Congress, the Internet Archive, in collaboration with the Library of Congress and WebArchivist.org, identified and archived Web sites associated with nearly 30,000 URLs. The purchase order that defined the scope of this task required that SUNYIT, in cooperation with the Library, select a group of approximately 2,500 URLs from the archive and, through the creation of metadata and a Web-based interface, create a collection identified as the "September 11 Web Archive."

Metadata Database Creation

The metadata database was to include a range of fields specified within a set of term definitions attached to project planning memorandum referenced in the purchase order (See Appendix A). SUNYIT was expected to create a plan for the database identifying the file structure and submit it to the Library for review. It was expected that the plan would describe the preferred technique to generate and store metadata and associate the metadata with base URLs. The base URLs were to be categorized by site producer, September 11 content, and Dublin Core elements to create the metadata database. The metadata database was to be provided as an XML file.

Web interface to collection

SUNYIT was expected to create a plan for a Web interface to the collection, and submit it to the Library for review. The Web interface was to be delivered, with documentation and installation instructions, and support provided for a total of 40 hours over 12 months following the Library's possession of the software. Technical support was to be made available at Contractor's convenience, by phone or email, and was to be designed to facilitate successful installation and operation of Web interface on the Library's enterprise IT system.

Final Report

SUNYIT was expected to provide a final report at the conclusion of the project, documenting the project expectations, results, and recommendations for future Web archiving projects.

Project Results

Acquisition and collection development

This project identified and cataloged 2,399 URLs for inclusion in the Library of Congress September 11 Web Archive. These URLs were selected from amongst those identified and archived between September 11, 2001 and November 30, 2001 by the Internet Archive. Under contract with the Library of Congress, the Internet Archive, in collaboration with the Library of Congress and WebArchivist.org, identified and archived Web sites associated with nearly 30,000 URLs. A purchase order defining the scope of this project required that SUNYIT, in cooperation with the Library, select a group of approximately 2,500 URLs from the archive and, through the creation of metadata and a Web-based interface, create a collection identified as the "September 11 Web Archive."

During the active collection process, between September 12, 2001 and December 1, 2001, staff affiliated with the Library of Congress identified 1,408 URLs for inclusion in the archive (See Appendix B). During this time, researchers and others associated with WebArchivist.org, including those visitors to a Web site maintained by WebArchivist.org soliciting URLs for inclusion in the collection, identified 29,367 URLs for inclusion in the archive (See Appendix C). A total of 28,560 distinct URLs were identified between these two processes (See Appendix D and Appendix E).
To identify URLs for consideration for inclusion in the archive, an analysis process that involved matching URLs against keywords was employed. This process selected 495 URLs that matched a set of keywords (See Appendix F) from among those identified by the WebArchivist project team. Combined with those URLs identified by the Library, an initial set of 1,903 URLs was presented to the Library. The initial list of URLs was reviewed by the Library, which conducted its own keyword search and other internal processes, and expanded to a list of 2,500 URLs. In the process of cataloging the URLs and reviewing the available archived objects, a number of duplicate and non-functioning URLs were identified. The final collection, including 2,399 records, was delivered to the Library as the September 11 Web Archive.

Appendix G provides a list of URLs and summary characteristics for each of the records included in the collection.

Metadata creation

In collaboration with the Library of Congress staff, researchers at WebArchivist.org developed a format for "Web Archive Records," building on work recently completed by SUNYIT for the Library on the Election 2002 Web Archive. For each record, it was determined that the fields, identified and described below in the "Web interface to collection" section, be included in the Web Archive records. The derivation of these fields is detailed in Appendix H.

WebArchivist researchers characterized the sites in the collection by producer type, following guidelines developed in cooperation with the Library of Congress. In addition, metadata for each site included in the archive was developed by post-processing data

collected by WebArchivist.org researchers. In addition to providing core descriptive information about sites in the archive, the metadata is used to populate the database that forms the basis of an interface to the archived sites. This interface is described in detail below.

All data in the metadata database was collected from the Web sites themselves. To represent the metadata elements, WebArchivist, following the direction of the Library, adopted the "Metadata Object Description Schema" (MODS), an XML schema developed by the Library of Congress' Network Development and MARC Standards Office. MODS enables the creation of original resource description records, and includes a subset of MARC fields and uses language-based tags. The standard is maintained by the Network Development and MARC Standards Office of the Library of Congress with input from users. An XML file containing data associated with all records in the collection was provided to the Library as part of the WebArchivist.org release of the September 11 Web Application (see discussion below in "Web interface to collection" section). Each site in the archive is represented by two elements – an "index" element and a "data" element – with a common identifier. The attributes in the index element are used as metadata to populate the search and search results aspects of the access interface, discussed below. All attributes in the index element were derived directly or indirectly from attributes in the data element. Attributes in the data element form the core metadata associated with each site included in the archive, and are included in the Web Archive records displayed by the interface, as discussed below. The attributes associated with both index and data elements are named and described in Appendix G of this report.

Web Interface to Collection

WebArchivist.org, in collaboration with the Library, developed a Web-based application providing access to the metadata and links to Web Archive Resource Pages associated with each site in the archive. The application links the various facets of the project together and provides a public face to the Web archive. Appendix I of this report provides illustrated screen shots of the application, which consists of two main components – Search and Web Archive Records.

Within the Search component, there are three separate subcomponents – Search Options, Search Path, and Search Results. Users are brought to an initial search page, with a configuration controlled by a set of configuration files; the discussion that follows is based on the configuration as delivered to the Library in April, 2004 (see discussion below). The search interface includes three primary subcomponents. The Search Options component, as illustrated in Appendix I: Figure 1, provides a list of categories within which users can select attributes of interest. As initially delivered, users could select attributes including first letter of producer name, producer type, producer country, language, bioterrorism content, September 11 content, or Afghan War content. Within each category, users could select from among any number of attributes. For example, within the producer type category, users could select from among 12 attributes – business, charity/civic, educational, ethnic, government, individual/volunteer, not catalogued, political party, portal, press, professional association/union, public interest/advocacy, or religious. Selecting a particular attribute in a category narrows the search to those records that share that value; the remaining categories and the attributes

within those categories then presented to the user (See Appendix I: Figure 2). The Search Path, presented across the top of the search interface records and highlighted in Appendix I: Figure 2, presents the path through the current search, and allows the user to return to a previous position in the search. The number of records available at each level of the search is also presented to the user. This innovative interface is especially appropriate in an archive of Web sites in which users may be searching for a specific group of sites that share a set of characteristics. For example, users interested in viewing the Web sites of political parties with September 11 content present could isolate the set of records that satisfy these conditions. The third subcomponent of the Search Interface, the Search Results section, is a listing of sites matching the selected attributes of the searched categories, as illustrated in Appendix I: Figure 2. Ten sites are presented on each page of search results, and a navigation menu to other pages is provided. The selected characteristics displayed are configurable in the application options. One of the characteristics, usually the name of the site, is designated as the linking characteristic, and provides a link to a Web Archive Record associated with the individual sites. The Search Results can be sorted by any of the characteristics by clicking on the name of the characteristic.

The second primary component of the Web application consists of Web Archive Records, created for each site in the archive. Appendix G includes the Web Archive Record URL for each of these records. Clicking on a site link within the Search Results returns the Web archive record for the selected site. The Web archive record, as illustrated in Appendix I: Figure 3, serves as the Library's front end to the archived sites. Each of these records includes an identical set of attributes:

- Abstract: provides a brief description of the contents of the web site
- Access condition: includes a statement of any restrictions placed on viewing the web site.
- Active site: the original URL of the site as archived.
- Alternate title: provides an alternate title for the site, in the event the title provided in the original HTML was not present or acceptable.
- Collection name: identifies each site as part of the "September 11 Web Archive."
- Dates captured: indicates the first and last dates for which impressions were archived for each site.
- Genre: identifies the media type represented by the archive item; in this project, all items were identified as "Web site."
- Identifier: is the URL of the site archived, and was used for identification purposes.
- Language: indicates the primary language of the Web site, identified following the practice of ISO 639-2 Bibliographic Code.

- Producer name: identifies the organization or individual represented by the website, using the form of name most commonly found.

- Subject: includes search headings using controlled keywords as specified by the Library of Congress.

- Title: extracted from the title tag on the Web site on the date closest to the election; if it was blank, the alternate title was substituted for the title.

Each Web Archive record also includes a link to the archived site, if the access conditions established for the site permit. The link to the archived site takes the user outside of the Web application developed by WebArchivist, and enters the user into the actual archive of Web sites. The Web Archive record is designed to serve as the "front door" to the archive of Web sites; scholars and others seeking to create links to archived sites within the collection are to be encouraged to create links to the enduring Web Archive record pages; the archived sites themselves may be reindexed or moved to other servers, thus rending deep links to the collection unstable.

The WebArchivist application, presented to the Library in final form in February 2004 as a compressed archive file in a tar format <minerva_911_20040227131543.tgz > and included as Appendix J of this report, was the primary deliverable of this agreement. The compressed archive file includes a primary data file, <911_data.xml> Modification to this file will change the data displayed in both the Search and Web archive record components.

Conclusions & Recommendations

Acquisition and collection development of rapidly emerging, unanticipated and unstable Web spheres

The notion of a Web sphere, first discussed in scholarly literature authored by the principal investigator of this report and a collaborator from the University of Washington, is a useful organizing principle for consideration of Web archives and collection.[1] A Web sphere can be conceptualized not simply as a collection of Web sites, but as a hyperlinked set of dynamically yet systematically defined digital resources spanning multiple Web sites deemed relevant or related to a central theme or "object." The Web sphere that developed around the events of September 11, 2001 posed significant challenges for acquisition and collection development because of its unanticipated and unstable nature and its rapid emergence.

Web spheres emerging quickly after an unanticipated event, as was the case with September 11, may be more difficult to acquire and collect, as a rapid investment of resources (e.g. time, money, topical expertise) may be required. In addition, the Web sphere emerging around September 11 was produced by both predictable and anticipated actors (for example, government agencies, relief and charity organizations, press organizations, and citizens) as well as less predictable set of actor types. Following the terrorist attacks of September 11, 2001, significant and unpredicted activity on web sites produced by corporations and businesses, along with more predictable activity on sites produced by religious organizations, educational institutions and government agencies, was observed. A predictable set of actors facilitates identification of a universe of sites to consider for inclusion in a collection; a less predictable set of actors makes this task more difficult, and requires additional searching and identification activities. Finally, the level of stability, referring to the frequency of entry and exit of new producers, the extent to which producers update, maintain and continue serving sites, and the frequency and breadth of changes to content and features within the web sites being considered for collection, may influence acquisition and collection decisions.

In this case, the decision was made to collect broadly, deeply and frequently. As a result, nearly 30,000 URLs were used as base URLs to seed a daily crawl by the Internet Archive. Given the extraordinary events, the Library was fortunate to establish a cooperative agreement with project collaborators, who began work on the project long before contracts were in place to cover the associated costs. Though this approach resulted in a very large collection of archived objects, and required the development of selection criteria to identify sites to be catalogued, this process undoubtedly resulted in

[1] See Schneider, S. M. and Foot, K. A. (forthcoming), "Web Sphere Analysis: An Approach to Studying Online Action," in Virtual Methods, Hine, C. (Ed., Berg Publishers, United Kingdom and Schneider, S. M. and Foot, K. A (2004), "The Web as an Object of Study," in New Media and Society 6:1, pp. 94-102.

the identification and collection of a very wide range of Web sites and pages that a more narrowly focused and less timely collection process would likely have missed. For example, a collection strategy emphasizing previously deployed Web sites, such as those produced by predictable actor types such as government agencies and charitable organizations, would have excluded some of the most interesting and innovative sites: those produced by individuals and volunteers, foreign governments and corporations.

In summary, one specific recommendation is made concerning the acquisition and collection development:

- Develop a strategic approach, including implementation of anticipatory contracts to acquisition agencies, to identifying and archiving Web sites emerging in unanticipated, unstable and unpredictable Web spheres, and be prepared to implement this strategy as events required.

Metadata creation

The metadata developed for the collection made possible the creation of an innovative Web application providing access to the archive. This metadata, as described above, is primarily focused on the characteristics of the site producers; limited attention is given to the characteristics of the Web sites. The one exception concerns the presence of content related to September 11, bioterrorism and the Afghan War.

One area of concern with respect to the collection of metadata is the notion of machine collection. The project anticipated collecting machine data for the title of the site, as well as the dates of collection. Analysis of the title data collected by machine indicates that such an approach is not as straightforward as first anticipated. A constructed title, based on metadata about the producer and developed by collection specialists, would have been more efficient and useful in the catalog process.

There are two specific problems associated with the machine collection of metadata. First, given the lack of standards applied by Web developers to even a basic HTML field such as title, useful data is often not accessible to machines. Second, and more importantly, the issue of non-standard characters and encoding, especially when translated to XML, causes significant problems with machine data. For the relatively few sites included in this archive, it would have been more efficient to use constructed title based on metadata developed by collection specialists.

In summary, one specific recommendation is made concerning the development of metadata:

- In relatively small collections (fewer than 5,000 catalogued objects), do not rely on machine collection of metadata for catalog fields.

Web interface to collection

The Web application developed for the Library to provide access to the collection and the metadata about the objects included in the collection is one of the most advanced and

fully-developed interfaces to a Web archive that has been developed to date. The interface offers an opportunity for the Library to provide search capabilities at the site level to a substantial collection of archived sites. This interface can serve as a model for future archiving projects should it be determined to meet the Library's needs.

The interface was developed largely as a stand-alone and single-project application. There was no expectation that the Library would receive a product that could be used, off the shelf, with other Web archives. Accordingly, few resources were devoted to developing the type of application that could be easily retooled for use with other archive collections. While this is certainly possible with the software as currently delivered, extension of the Web application to other environments requires technical skills and expertise.

Absent any specifications from the Library concerning the nature of the server on which the application was to reside, the interface was designed to run in an environment with low processing demands. This restricted the development of server-intensive searching capabilities. Further, the Library indicated that full-fledged searching capabilities would be better handled by other projects in development.

Future archiving projects could benefit from additional specifications for the Web application. In summary, a single recommendation is made with respect to the Web application:

- Include detailed specifications for Web application to serve as interface to archive in overall project plan.

Appendices

Appendix A: Definitions and common elements, as developed for Project Scope Document

"Analysis Set" means the group of approximately 2,500 Base URLs selected by Contractor from Archive using Process.

"Archival URL" means a URL used for retrieving Archived Objects from the Archive, as specified in the Internet Archive's documentation.

"Archive" means the set of ARC files collected by the Internet Archive on behalf of the Library of Congress, from September 11, 2001 to December 1, 2001, in response to requests from the Library of Congress and WebArchivist.org; and the set of DAT files and CDX files generated by the Internet Archive using the ARC files.

"Archive Access" means the ability of Contractor to both query the Archive via a shell account maintained by the Library on the Internet Archive servers or other servers containing the Archive, and to use the Wayback Machine to query the Archive; and the ability to view archived pages using an archival URL format via a supported Web browser over the Internet.

"Archived Object" means a file contained in the archive accessible with an Archival URL.

"Base URL" means a web page sent to a web browser as a result of a query submitted to the Wayback Machine in Archival URL format.

"Base URL page"means a web page sent to a web browser as a result of a query submitted to the Wayback Machine in archival URL format.

"Base URL" means a URL that is an element in the Requested Set or the Analysis Set.

"Contractor" means the Research Foundation of the State University of New York and its subcontractors.

"Dublin Core Elements" means the set of elements to be categorized by Base URL. When possible, data for these elements will be captured from the text or metadata embedded within the HTML source file of Base URL pages. These elements include the following:

Title: Text included within the title tag of the HTML source file of a Base URL page.

Author or Creator: Name of the entity who appears to be primarily responsible for making the content of a Base URL page, as identified by reference to text or graphics on a Base URL page, or by reference to an about us page linked from a Base URL page.

Description: A brief one-sentence description of the site associated with a Base URL page, shall be generated, referencing a possible site producer and identifying a possible purpose of the site.

Publisher: The entity who appears to be responsible for making a Base URL page available, as identified by reference to text or graphics on a Base URL page, or by reference to an about us page linked from a Base URL page.

Date: The archived time associated with the Base URL page archived closest to and after 9:00 AM EST on September 11, 2001, using the form YYYYMMDDHHMMSS, such that the date 20010911090000 corresponds to 9:00:00 AM on September 11, 2001

Resource Type: Each URL shall be identified as a "web site."

Format: A list of the distinct file formats of all archived objects associated with a Base URL page.

Resource Identifier: Base URL

Language: The primary language of a Base URL page shall be identified following the practice defined in RFC 1766.

Coverage: The amount of time, in seconds, between the archived time associated with the Base URL page archived closest to and after 9:00 AM EST on September 11, 2001, and the archived time of the Base URL page archived closest to and before 11:59:59 PM EST, December 1, 2001.

Rights Management: An identifier, provided by the Library of Congress, associated with a Base URL.

"Internet Archive" is the organization that has contracted with the Library to collect, archive and index web pages, and to provide the Wayback Machine.

"Library" means the Library of Congress.

"Metadata Database" means a XML document containing data related to Archived Objects associated with Base URLs in the analysis set, including the following entities and attributes: Base URL, Site Producer, September 11 Content, and the Dublin Core Elements.

"Process" means the steps taken to select the Analysis Set from the list of Base URLs, provided by the Library and others, to the Internet Archive between September 11, 2001 and December 1, 2001, for inclusion in the Archive. The steps include: 1) Analyze URLs to determine the estimated number of Base URL pages in the Archive, and the time those pages were archived. 2) URLs will be selected in to favor URLs with pages collected closest to September 11, and reflects the editorial judgment of the Contractor.

"Review" means the Library's consideration and approval of plans the Contractor submits to the Library. For plans the Contractor submits, the Library shall have ten (10) business days to review and approve such plans. The absence of such comments shall indicate concurrence. If the Library suggests revisions, the Contractor shall consider the suggested revisions, and within ten (10) business days of receipt of the revisions, submit

a revised plan for the Library to Review. This Review process shall continue until the Library and the Contractor agree on the plan.

"September 11 Content" means the presence of text or graphics reflecting the events of September 11 or its aftermath on at least one of three Base URL Pages, selected from among those available in the archive, distributed as evenly as possible along a range of dates between September 11, 2001 and October 28, 2001.

"Site Producer" means the type of entity who appears to have published or produced a Base URL page. Categories will include media/press; government; individual/volunteer; charity/religious; commercial; corporate, with the addition of other categories as necessary.

"Wayback Machine" means a HTTP interface for querying the set of Archived Objects maintained by the Internet Archive, responding to requests from Internet Explorer Version 5.5, and Netscape Navigator Version 5.0. for Archival URLs.

"Website Software" means software to facilitate access via Internet Explorer Version 5.5 and Netscape Navigator Version 5.0 to Base URLs in the Analysis Set, and to Base URL pages in the Archive, using the Metadata Database, the Wayback Machine, a relational database, a Tomcat Servlet Engine and Java. The software will allow (1) searching based on all fields included in the Metadata Database, with the results of the searches identifying Base URLs matching the query, as well as the number and date range of Base URL pages matching the query; (2) randomly selecting and visiting a Base URL page; (3) selecting and visiting a random Base URL page matching specified criteria selected from among all fields included in the Metadata Database; (4) searching for a Base URL matching text fragments found on a Base URL page; and (5) generating a list of Base URLs matching specified criteria based on all fields included in the Metadata Database.

Appendix B: URLs submitted by Library of Congress to Internet Archive for collection (Overview of electronic appendix)

File Name: loc_sites.txt

Description of file: This file includes all URLs submitted by the Library of Congress to the Internet Archive for collection between September 12, 2001 and November 30, 2001. The file contains one url per line. A total of 1,408 entries are included in the file.

Appendix C: URLs submitted by WebArchivist.org to Internet Archive for collection (Overview of electronic appendix)

File Name: wa_sites.txt

Description of file: This file includes all URLs submitted by WebArchivist.org to the Internet Archive for collection between September 12, 2001 and November 30, 2001. The file contains one url per line. A total of 29,367 entries are included in the file.

Appendix D: Distinct URLs submitted by Library of Congress to Internet Archive for collection (Overview of electronic appendix)

File Name: loc_distinct_sites.txt (TXT format)

Description of file: This file includes all URLs submitted by the Library of Congress to the Internet Archive for collection between September 12, 2001 and November 30, 2001 not also submitted by WebArchivist.org. The file contains one url per line. A total of 601 entries are included in the file.

Appendix E: Distinct URLs submitted by WebArchivist.org to Internet Archive for collection (Overview of electronic appendix)

File Name: wa_distinct_sites.txt

Description of file: This file includes all URLs submitted by WebArchivist.org to the Internet Archive for collection between September 12, 2001 and November 30, 2001 not also submitted by the Library of Congress. The file contains one url per line. A total of 28,560 entries are included in the file.

Appendix F: Keywords used to identify URLs for consideration for inclusion in the September 11 Web Archive

911
afgan
america
anthrax
arab
attack
bio
bomb
bush
chemical
devil
disaster
explosion
god
justice
laden
muslim
newyork
nuclear
osama
pentagon
relig
satan
sep11
sept11
september11
septembereleven
terror
tragedy
united
war
worldtrade
wtc

Appendix G: Sites included in September 11 Web Archive

File Name: sites-in-archive.html (HTML format)

Description of file: This file provides a list of sites included in the September 11 Web Archive. For each site, the following data is provided:

Column	Heading	Description
1	Producer Name	Name of producer, as included in data file
2	Site ID #	Unique identifer; use to find Web Archive Record
3	First Capture Date	Date of first impression of site in archive
4	Last Capture Date	Date of last impression of site in archive
5	URL	Original URL of site as captured

Appendix H: Derivation and Description of Fields in Metadata Database

Field Name (XML)	Description
abstract	Derived content, following format: [siteName], a Web Site produced by [abstractProducerPresentation], is part of the Library of Congress September 11 Web Archive and preserves the web expressions of individuals, groups, the press and institutions in the United States and from around the world in the aftermath of the attacks in the United States on September 11, 2001.
abstractContentPresentation	Summary of content (not used)
abstractProducerPresentation	[constructedProducerName], (derived [producerType]
accessCondition	"Statement to be provided by Library of Congress" (constant)
afghanWarContent	Presence of content related to bioterrorism on at least one of three selected impressions of archived site (first impression, middle impression, last impression)
altTitle	Alternative title, based on site producer name
bioterrorContent	Presence of content related to bioterrorism on at least one of three selected impressions of archived site (first impression, middle impression, last impression)
collection	September 11 Web Archive (Constant)
collectionNumber	1 (Constant)
constructedProducerName	Derived field used for Web Archive record
endCaptureDate	Latest date associated with site in archive; determined by reference to archive index files (not verified with reference to availability or viewability of actual impression)
id	Identification number assigned to URL
identifier	Base URL
language	The primary language of the Web site.
nameType	Same as producerNameType, not used
producerCountry	The country of the site producer
producerName	Name of producer, as identified by reference to "about us" or "contacts" page on impression of first viewable impression of site.
producerNameCorporate	Corporate name of site producer
producerNameType	Type of producer, either Corporate or Individual, following guidelines provided by Library of Congress
producerType	Type of producer, either business, charity/civic, educational, ethnic, government, individual/volunteer, not catalogued, political party, portal, press, professional association/union, public interest/advocacy, or religious group, following guidelines provided by the Library of Congress
september11Content	Presence of content related to September 11 terrorist attacks on at least one of three selected impressions of archived site (first impression, middle impression, last impression)
siteName	Name of site, as identified by Webarchivist.org researchers
startCaptureDate	Earliest date associated with site in archive; determined by

Field Name (XML)	Description
	reference to archive index files (not verified with reference to availability or viewability of actual impression)
subjects	Subjects, as specified by Library of Congress, dependent on content. If content related to September 11 was present, subject tags included <subject><topic>September 11 Terrorist Attacks, 2001</topic></subject>. If content related to Afghan War was present, subject tags included <subject><geographic>Afghanistan</geographic><topic>History </topic><temporal>2001-</temporal></subject>. If content related to Bioterrorism was present, subject tags included <subject><topic>Bioterrorism</topic></subject>.
title	The title of the Web site, based on title tag on impression captured closest to and after 9:00 AM EST on September 11, 2001.
yearCollected	2001 (Constant)

Appendix I: Illustrated Screenshots of Web Interface

Figure 1: Screen Shot of Opening Page to September 11 Web Archive

Figure 2: Screen shot of Web interface to September 11 Web Archive highlighting search path and search results

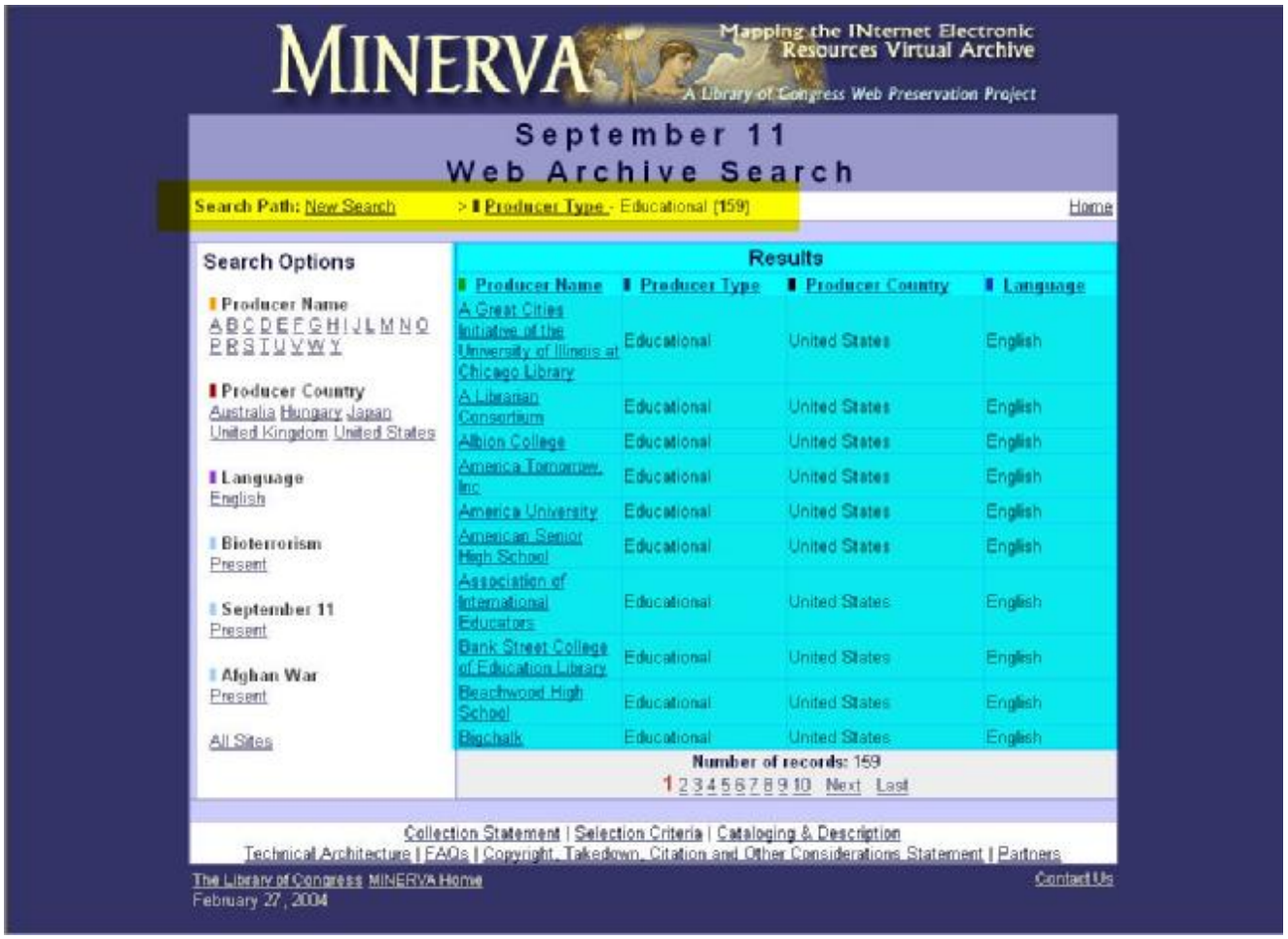
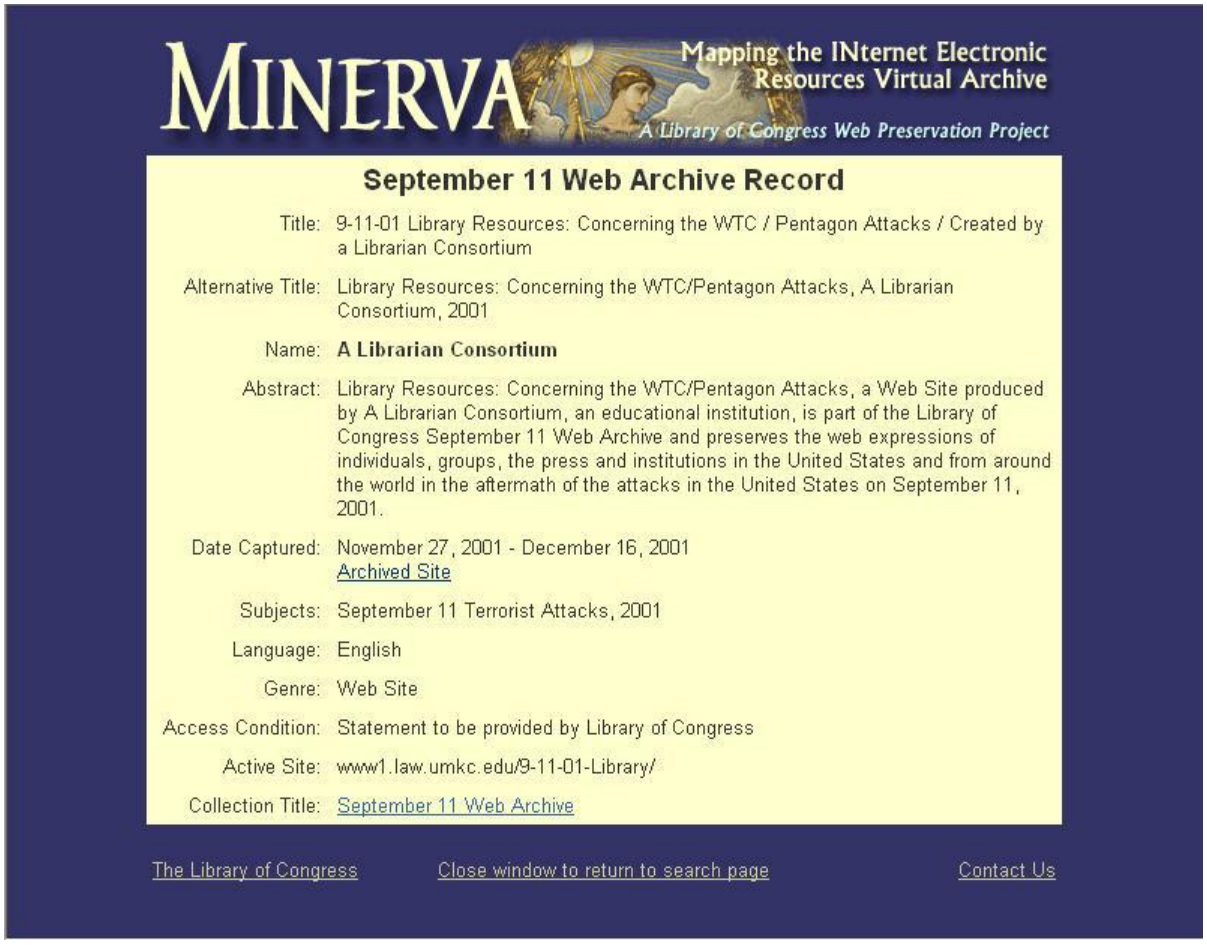


Figure 3: Screen shot of Web Archive Record component of September 11 Web Archive



Appendix J: Compressed Archive of Web Application (Overview of Electronic Appendix)

File Names: minerva_911_20040227131543.tgz (tar gzip format), Minerva-docs.xm. (xml format)

Description of files: These file provide the delivery and documentation of the September 11 Web Archive, including the data files. The file <Minerva-doc.xml> provides the documentation and should be consulted first. The gzip file can be opened using standard zip extractors, and includes a single tar archive, which can be extracted using standard zip extractors, including WinZip and other similar products.